CLAIMS:

What is claimed is:

1. A method for determining the impact and influence of data cleaning operations into the results of data mining analysis comprising the steps of:

generating a set of cleaning attributes for each cleaned data record in a complete set of cleaned data records, said cleaning attributes reflecting which fields of each record have been modified by a cleaning operation;

receiving a data feature identified by a data mining process for a subset of said complete set of cleaned data records;

determining a degree of correlation of said data feature to the modified fields of said subset of cleaned data records according to said cleaning attributes; and

declaring said data feature as suspect responsive to said degree of correlation exceeding a threshold.

2. The method as set forth in Claim 1 wherein said step of generating a set of cleaning attributes comprises generating a set of bit-mapped Boolean flags to form a cleaning attributes register for each cleaned data record.

3. The method as set forth in Claim 1 wherein said step of generating a set of cleaning attributes comprises performing an operation selected from the group of appending a set of cleaning attributes to each cleaned data record,

prepending a set of cleaning attributes to each cleaned data record, distributing a set of cleaning attributes to each cleaned data record, and generating a cleaning attribute table.

4. The method as set forth in Claim 1 wherein said step of receiving a data feature comprises a step selected from the group of receiving a cluster, receiving a trend, and receiving a pattern.

5. The method as set forth in Claim 1 wherein said step of generating a set of cleaning attributes for each cleaned data record in a complete set of cleaned data records comprises comparing each record in a raw data set to each record in a cleaned data set.

6. A data structure comprising:

one or more data records, each record having a plurality of data fields; a set of cleaning attributes for each data field in each data record indicating which fields have been modified by a data cleaning operation; and a means for associating said cleaning attributes with said data fields.

7. The data structure as set forth in Claim 6 wherein said cleaning attributes comprise Boolean flags.

8. The data structure as set forth in Claim 6 wherein said data records comprise rows in a cleaned data table, wherein said set of cleaning attributes comprise subsets in a cleaning attributes table, and wherein said means for associating said cleaning attributes with said data fields comprises a row index.

9. The data structure as set forth in Claim 6 wherein said data records comprise records in a database, wherein said set of cleaning attributes comprise subsets in a cleaning attributes contained in said records, and wherein said means for associating said cleaning attributes with said data fields comprises a means selected from the group of appending, prepending and distributing said cleaning attributes in each record.

10. A computer readable medium encoded with software for determining the impact and influence of data cleaning operations into the results of data mining analysis, said software performing the steps of:

generating a set of cleaning attributes for each cleaned data record in a complete set of cleaned data records, said cleaning attributes reflecting which fields of each record have been modified by a cleaning operation;

receiving a data feature identified by a data mining process for a subset of said complete set of cleaned data records;

determining a degree of correlation of said data feature to the modified fields of said subset of cleaned data records according to said cleaning attributes; and

declaring said data feature as suspect responsive to said degree of correlation exceeding a threshold.

11. The computer readable medium as set forth in Claim 10 wherein said software for generating a set of cleaning attributes comprises software for generating a set of bit-mapped Boolean flags to form a cleaning attributes register for each cleaned data record.

12. The computer readable medium as set forth in Claim 10 wherein said software for generating a set of cleaning attributes comprises software for performing an operation selected from the group of appending a set of cleaning attributes to each cleaned data record, prepending a set of cleaning attributes to each cleaned data record, distributing a set of cleaning attributes to each cleaned data record, and generating a cleaning attribute table.

13. The computer readable medium as set forth in Claim 10 wherein said software for receiving a data feature comprises software for performing a step selected from the group of receiving a cluster, receiving a trend, and receiving a pattern.

14. The computer readable medium as set forth in Claim 10 wherein said software for generating a set of cleaning attributes for each cleaned data record in a complete set of cleaned data records comprises software for comparing each record in a raw data set to each record in a cleaned data set.

15. A system for determining the impact and influence of data cleaning operations into the results of data mining analysis, comprising:

a set of cleaning attributes for each cleaned data record in a complete set of cleaned data records, said cleaning attributes reflecting which fields of each record have been modified by a cleaning operation;

a data feature received from a data mining process for a subset of said complete set of cleaned data records;

an analyzer for determining a degree of correlation of said data feature to the modified fields of said subset of cleaned data records according to said cleaning attributes; and

a reporter for declaring said data feature as suspect responsive to said degree of correlation exceeding a threshold.

16. The system as set forth in Claim 15 wherein said set of cleaning attributes comprises a set of bit-mapped Boolean flags which form a cleaning attributes register for each cleaned data record.

17. The system as set forth in Claim 15 wherein said a set of cleaning attributes are associated with said cleaned data records using an association method selected from the group of appending a set of cleaning attributes to each cleaned data record, prepending a set of cleaning attributes to each cleaned data record, distributing a set of cleaning attributes to each cleaned data record, and generating a cleaning attribute table.

18. The system as set forth in Claim 15 wherein said received data feature comprises a data feature selected from the group of a cluster, a trend, and a pattern.